Guidelines for reporting protein modelling studies

Mauno Vihinen[1]

[1]Lund University

May 7, 2021

Abstract

Computational modelling tools are widely used, however, articles describing modelling studies frequently do not contain sufficient details to allow the reader to comprehend the modelling procedure, quality of the produced model and validity of interpretations and predictions made based on the model. Here, guidelines were developed for items that have to be included when reporting studies and results based on protein modelling. A brief and concise checklist of required data items was compiled. These guidelines are simple to follow and apply, but require meticulous description of details, many of which can be placed to supplementary material. Authors have to pay attention to details when reporting modelling process. The generated structural models should be made publicly available, preferably by submitting to one of the existing repositories.

Introduction

Three dimensional structures of proteins and other macromolecules are often required for detailed studies and for understanding of mechanisms and functions. Although new structures are determined at fast pace, the gap between known sequences and structures is widening (Kc, 2017). Due to low cost and availability of sequencing facilities, numbers of genes and genomes of novel organisms are expanding. Structures are still missing for many proteins even in well-studied organisms. For example, structures for a large part of human proteome are not yet available. Protein modelling may be a useful alternative when experimental structure does not exist. Models and the process of predicting them is often poorly documented in scientific literature. Unfortunately, this applies to descriptions of experimental studies as well. Analysis of 268 biomedical publications with experimental data revealed that only one article (0.37%) included full report of protocols (Iqbal et al., 2016). This paper provides guidelines for detailed and proper reporting of modelling studies.

Protein structures can be modelled by using methods from three major categories of tools, for a review see (Dorn et al., 2014; Kuhlman and Bradley, 2019). *Ab initio* methods predict structures from scratch based on the most favorable energy conformations. These methods require extensive computational resources. The goal for fold recognition (also called threading) methods is to reveal the folding type of the protein of interest. Homology modelling based on a related, known structure or several ones provides often the most reliable atom level models. All these approaches are complicated and include several steps. Full description of these studies is an exception rather than a norm. Inadequate descriptions prevent full comprehension of the models, investigating their quality and evaluation, extension and repetition to new studies. Proper reporting would allow readers to pick problematic cases and details if peer-review has failed in detecting the deficiencies. In recent years scientific communities have awakened to reproducibility (repeatability) crisis (Baker, 2016; Begley and Ioannidis, 2015), one reason for which is inadequate description of studies.

Strategies and tips have been published for how to model structure (Dhingra et al., 2020; Dorn et al., 2014; Haddad et al., 2020), however, there are not instructions for description of protein models despite a recognized need (see e.g. (Schwede et al., 2009)). The only available guidelines are for small molecules related to medicinal chemistry, mainly drugs (Gund et al., 1988). Related recommendations and guidelines have been published in bioinformatics e.g. for computer-aided variation interpretation (Vihinen, 2012, 2013) and for sequence alignments (Vihinen, 2020). Minimum Information About Bioinformatics Investigation (MIABI, https://fairsharing.org/FAIRsharing.28yec8) (Tan et al., 2010) provides basic reporting guidelines including the used algorithm, analysis protocol, used databases, resources, software and (web)services. However, more detailed data is needed for structural models.

The goal in here is to provide guidelines for reporting use and results from molecular modelling. The guidelines originate from frustration in reading published articles and manuscripts that do not contain sufficient details to allow reader to comprehend and evaluate modelling process and quality of the output and validity of made predictions e.g. in relation to function or diseases. These guidelines are simple to follow and apply. Modelling studies should be described with similar degree of detail as experimental articles. As many journals do not allow reporting full methodological details in the main article, provide details in supplementary material and other parts of the article. Many details can be included also to tables, figures and figure captions.

Guidelines for protein modelling

A simple checklist of items necessary to describe computer modelling studies is presented. Each of the items is provided with a brief description. Which items of the guidelines should be included in a certain article depends on the used modelling strategy. Before applying any computational analyses, one has to know the quality of used data as computational predictions can be useless if the starting point is wrong or severely biased.

Precalculated models for many proteins are available at ModBase (Pieper et al., 2014) and SwissModel Repository (Bienert et al., 2017), however, they do not provide full details as instructed in here. These models are predicted with automated pipelines, thus the modelling procedure is not optimized for each structure. Such models can still be useful for various applications when experimental structures have not been determined. When close enough structures are available, automatic models may be of higher quality than an inexperienced modeler can achieve. Users of any types of structural predictions must bear in mind and be aware of limitations of such models.

1. Describe purpose and objective of the modelling exercise

The purpose and goal of the modelling experiment should be provided as it affects the choice of the approach as well as what types of biological predictions can be made. Model quality and accuracy depends on the available data and modelling process. For example, detailed intra- or intermolecular contacts can only be investigated with homology modelling. On the other hand, fold recognition methods may be sufficient for detecting secondary structural elements and their organization.

Example: The structure of the Bruton tyrosine kinase (BTK) pleckstrin homology (PH) domain was modelled to facilitate structure-based interpretation of disease-causing variants in X-linked agammaglobulinemia (XLA).

2. Use systematic names and descriptions

Systematic descriptions should be used in relation to models (and elsewhere), when systematics exists. Relevant areas in relation to molecular models include names of organisms (taxonomy), gene and protein names (Gray et al., 2015), variations in proteins (HGVS nomenclature) (den Dunnen and Antonarakis, 2001), names of files (sequence and PDB ids), variation effects and consequences (Variation Ontology annotations (Vihinen, 2014)), etc. In the case of sequence ids, version numbers must be included unless LRGs (Dalgleish et al., 2010) are used.

Example: Variations in human *BTK* gene and translated protein (LRG_128) lead to XLA, a X-chromosomal primary immunodeficiency (OMIM # 300755). Variant p.R28C has been identified in XLA patients.

3. Justify choice the modelled protein

Think carefully which protein (or proteins) is the most relevant to model to address your biological questions. If you are interested in just a certain protein the choice is simple. In the case of networks or e.g. complexes there may be several options to consider. Describe why the studied protein was selected.

Example: BTK PH domain was modelled since the protein is medically important and it contains numerous disease-related variants structural mechanistic bases of which are unknown.

4. Justify choice of template(s)

For homology modelling, the choice of the template(s) is crucial. Sequence similarity is an important factor, but there are also other issues. If several different types of structures are available, then a choice of e.g. apo vs holo form or active vs inactive conformation has be be described. Many proteins have several conformations, functional relevance of which varies. If you are interested on a specific part or domain within a large protein, it obviously should be included to the template. In case of drugs and other ligands, it may be more relevant to have a template with similar or related compound in relevant protein conformation than to take a structure with best resolution, R factor or other crystallographic or NMR quality measures. Mention PDB ids of the templates.

Example: The structure was modelled by using 9xyz as a template. The template protein shares the highest sequence identity with the protein (49%) among existing structures. There is a minimal number of gaps and they are short (2 gaps of two and three residues, respectively) and the sequences match along the entire chain length. Further, the investigated protein and the template have close substrate specificity.

5. Describe quality of template(s)

Higher resolution structures are preferred if a choice can be made between otherwise equally relevant structures. In the case of NMR structures, use stereochemical and other quality estimates. The regions most important for the intended use of the model should be reliably determined. Although the descriptions of templates can be found from articles and databases when using published structures, describe quality measures relevant for your choice and model application. If template is not published, even more specific descriptions are needed. In that case, include also the coordinates and description of the structure determination.

Sequence similarity and identity between the template and the protein to be modelled protein sets limits for quality of the model. If several related structures from different organism or for paralogs are available, they can provide additional evidence for the modelled structure.

Example. The template structure has been determined on high resolution (1.45 Å) and it shows good quality parameters (e.g. R_{free} is 0.195). B values are low (<30) for the functionally relevant substrate binding region and catalytic site and secondary structural elements in the protein core. PROCHECK (Laskowski et al., 1993) report did not indicate any quality issues.

6. Provide detailed sequence alignment information

Sequence alignment is instrumental for many modelling applications. Full details of the alignments must be included. These include used method, version, substitution table and program parameters. Notes like "default parameters" are not sufficient as they may differ between program installations and versions. Multiple sequence alignments can provide more reliable results than pairwise analyses when working with less conserved sequences. If any manual interventions have been made, they must be described and justified.

Follow guidelines for describing sequence alignments (Vihinen, 2020). Include database identifiers for sequences.

Example: The multiple sequence alignment of TEC family members included entries P51813 for BMX, LRG_128 for BTK, Q08881-1 for ITK, P42680-1 for TEC, and P42681-1 for TXK. The alignment was performed on Clustal Omega program (Clustal O(1.2.4)) (Sievers et al., 2011) and run at https://www.ebi.ac.uk/Tools/msa/clustalo/. The used substitution matrix was of Gonnet et al. (Gonnet et al., 1994). The program parameters were: Output guide tree, false; Output distance matrix, false; Dealign input sequences, false; mBed-like clustering guide tree, true; mBed-like clustering iteration, true; Number of iterations, 0; Maximum guide tree iterations, -1; Maximum HMM iterations, -1; Output alignment format, clustal_num; Output order, aligned; Sequence type, protein.

The insertion between residues 34 and 38 was manually adjusted so that there was just one gap instead of two provided by the program. The alignment covers 96.4-100% of the sequence lengths. The multiple sequence alignment is in Supplementary material.

7. Provide detailed description for model building

All the steps of the modelling procedure should be provided. For that purpose, provide information about the program(s) used, including version number. All the options and parameters used to adjust the program have to be included. Describe how amino acid substitutions were implemented. If loops have been modelled, provide information for algorithm, databases and program parameters. The approach used for building loops can have substantial effect on model quality and applicability.

If any manual steps are included, describe how they have been implemented and justify the choices made.

Example: The structure was modelled with program MODELLER (10.0, released Feb. 2nd, 2021 (Martí-Renom et al., 2000)) using UCSF Chimera (1.15) (Pettersen et al., 2004) interface. The Dunbrack 2010 rotamer library (Shapovalov and Dunbrack, 2011) was used to model substitutions to fit the sidechains into the structure with typical rotamer angles. Deletions and insertions were modelled based on the sequence alignment by replacing an existing loop or connecting segment. DOPE - Discrete Optimized Protein Energy score (Shen and Sali, 2006) with Lennard-Jones potential and GB/SA implicit solvent interaction was used for loop building.

8. Provide detailed refinement information

In this section, the applied methods and algorithms as well as parameters, options and protocols within them have to be described along with the justifications for the choices made. It is important to explain how extensive refinement has been made, including force fields, time steps, calculation details, stopping criterion, etc. whatever is relevant for the used method.

Example: The model was refined by energy minimization with the program CHARMM (version 41, release c40b2) (Brooks et al., 2009) using the all-hydrogen parameter set 22 in a stepwise manner whereby the conserved regions were constrained and the largest deviations were on the insertions/deletions. First the new loops were minimized. After 1000 cycles, the insertion/deletion regions were subjected to simulated annealing from 2000 to 300 K, followed by minimization for 500 cycles. Then the C_α atoms of the conserved regions, ATP, and Mg atoms were harmonically constrained, and the structure was minimized until the rms gradient was below 0.001 kcal/(mol·Å).

9. Validate and estimate model quality

Evaluate and estimate the quality of the structure. You can use the same tools as for experimental structures. Ramachandran plot can be investigated and analyzed with several tools. PROCHECK (Laskowski et al., 1993), WhatCheck (Hooft et al., 1996), MolProbity (Williams et al., 2018), Verify3D (Lüthy et al., 1992) and many other tools are available. Several of them can be run via SAVES (https://saves.mbi.ucla.edu/) to control stereochemical quality, bond angles, bond lengths, etc. At the Critical Assessment of protein Structure Prediction (CASP) challenges (https://www.predictioncenter.org/) the community has developed various tools and measures e.g. to compare structures with experimental structures.

One way to estimate usefulness of the model is to use it to interpret experimental data. If successful, the model can be useful also for other applications. Describe any details of such applications.

Example. The quality of the produced model was evaluated with PROCHECK (Laskowski et al., 1993) and MolProbity (Williams et al., 2018). Ramachandran plot (Ramachandran et al., 1963) indicated that 91% of the residues are on the most favored area and only two in disallowed region. Bond lengths and angles are well within normal ranges (on average 0.011Å and 2.1 degrees, respectively), the same with the torsion angles. According to MolProbity there are no clashes and there are not issues with geometry. The model was tested also by the Verify3D technique (Lüthy et al., 1992) and found to have the score of a typical globular protein.

As as functional test we evaluated the method in describing a known interaction based on a amino acid substitution. D304L in the lower lobe, was explained based on the model as having a structural effect originating from the disruption of a salt bridge with R244 in the upper lobe. The prediction is correct based on experimental data.

10. Consider limitations of the model

It is important to bear in mind that models are predictions and therefore have their limitations. This has to be considered in interpretation and explanation of e.g. in structure-based biological and medical phenomena. Describe any limitations of the generated model and limitations in the interpretation and inference based on it.

Example: The model is considered to be rather reliable, as indicated by validation scores. Since the template matches along with the entire sequence and has 53% sequence identity, the scaffolding and conserved parts are likely of high quality. The orientation of loop at position 216-220 is somewhat unsure. This region has relatively high B factor values in the template (>60 Å^2) and the additional residues on the surface loop have plenty of degrees of freedom. The structure was minimized to solve clashes and non-favorable stereochemistry.

11. Share the model

The produced model(s) has to be shared so that others can use, evaluate and improve it. PDB is the standard file format for transferring structures, however, there are also some other rather widely used formats. ModelArchive (https://modelarchive.org/) and BioStudies (https://www.ebi.ac.uk/biostudies/) (Sarkans et al., 2018) are repositories that share structure models and should be favored instead of own/departmental web sites.

Example: The model is available in BioStudies database (https://www.ebi.ac.uk/biostudies), accession number XYZ.

12. Interpret biological phenomena

With the generated model various biological and medicine-related aspects could be explained. Based on type of model, various aspects related to the structure, protein function, regulation, modification, activation etc. can be predicted. In addition to several CASP challenges also Critical Assessment of protein Function Annotation (CAFA, https://www.biofunctionprediction.org/cafa/) challenges have addressed and boosted development of methods for biological interpretation. As an example, a study describes in addition to the quality of model, relevance and accuracy of numerous biological and medical predictions made based on 15 blind prediction models (Khan and Vihinen, 2009).

Example: The conserved R520 is located at the N terminus of the catalytic loop where it is structurally and functionally vital. In cAPK, the corresponding R165 interacts with the stable T197 phosphorylation site. Both of these residues are conserved also in BTK; however, the corresponding phosphorylated residue is tyrosine in PTKs, Y551 in BTK. This residue precedes the homologue for T197 in cAPK. Many kinases are activated by phosphorylation at either of these sites. According to the BTK model, the phosphorylated Y551 could interact with R520. Variation R520E causes XLA without detectable B lymphocytes, presumably because of the lack of contact between the catalytic loop and the regulatory phosphorylation site.

13. Present clear and concise visualizations

Visualizations are essential and critical for describing biological, medical and other interpretations made based on the generated model. Instructions and tips have been published for powerful and clear visualizations, see (Johnson and Hertig, 2014; Mura et al., 2010), these should be followed. Keep figures simple and concentrate on the main message. Use rather several panels instead of trying to put all details in a single presentation. It is important to describe all pertinent aspects of the figures in the legends.

Example: Examples of informative visualizations along with instructions how to prepare them are provided in the articles (Johnson and Hertig, 2014; Mura et al., 2010).

Summary

A checklist of items to include in publications describing protein modelling and use of such models was developed. The presented guidelines are straightforward to follow and implement. However, authors have to be meticulous to record all necessary details. By doing so, the produced data will facilitate readers to evaluate, validate and use the generated models for further purposes. Models can be valuable and take substantial effort to generate. It is important to share them, preferably at one of the available repositories.

These guidelines are applicable also to other structural modelling-related studies, such as docking, binding and protein-protein interaction predictions (Aderinwale et al., 2020). Technical and other details and requirements discussed above are essential for understanding and evaluation of these studies, as well.

Conflict of interest disclosure

The authors declare no competing interests.

Funding statement

Financial support from Vetenskapsrådet (2019-01403) and Swedish Cancer Society (CAN 2017/699) is gratefully acknowledged.

References

Aderinwale, T., Christoffer, C.W., Sarkar, D., Alnabati, E., and Kihara, D. (2020). Computational structure modeling for diverse categories of macromolecular interactions. Curr Opin Struct Biol *64* , 1-8.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. Nature *533* , 452-454.

Begley, C.G., and Ioannidis, J.P. (2015). Reproducibility in science: improving the standard for basic and preclinical research. Circ Res*116* , 116-126.

Bienert, S., Waterhouse, A., de Beer, T.A., Tauriello, G., Studer, G., Bordoli, L., and Schwede, T. (2017). The SWISS-MODEL Repository-new features and functionality. Nucleic acids research *45* , D313-d319.

Brooks, B.R., Brooks, C.L., 3rd, Mackerell, A.D., Jr., Nilsson, L., Petrella, R.J., Roux, B., Won, Y., Archontis, G., Bartels, C., Boresch, S.*, et al.* (2009). CHARMM: the biomolecular simulation program. J Comput Chem *30* , 1545-1614.

Dalgleish, R., Flicek, P., Cunningham, F., Astashyn, A., Tully, R.E., Proctor, G., Chen, Y., McLaren, W.M., Larsson, P., Vaughan, B.W.*, et al.* (2010). Locus Reference Genomic sequences: an improved basis for describing human DNA variants. Genome Med *2* , 24.

den Dunnen, J.T., and Antonarakis, S.E. (2001). Nomenclature for the description of human sequence variations. Hum Genet *109* , 121-124.

Dhingra, S., Sowdhamini, R., Cadet, F., and Offmann, B. (2020). A glance into the evolution of template-free protein structure prediction methodologies. Biochimie *175* , 85-92.

Dorn, M., MB, E.S., Buriol, L.S., and Lamb, L.C. (2014). Three-dimensional protein structure prediction: Methods and computational strategies. Comput Biol Chem *53pb* , 251-276.

Gonnet, G.H., Cohen, M.A., and Benner, S.A. (1994). Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix. Biochem Biophys Res Commun *199* , 489-496.

Gray, K.A., Yates, B., Seal, R.L., Wright, M.W., and Bruford, E.A. (2015). Genenames.org: the HGNC resources in 2015. Nucleic acids research *43* , D1079-1085.

Gund, P., Barry, D.C., Blaney, J.M., and Cohen, N.C. (1988). Guidelines for publications in molecular modeling related to medicinal chemistry. J Med Chem *31* , 2230-2234.

Haddad, Y., Adam, V., and Heger, Z. (2020). Ten quick tips for homology modeling of high-resolution protein 3D structures. PLoS computational biology *16* , e1007449.

Hooft, R.W., Vriend, G., Sander, C., and Abola, E.E. (1996). Errors in protein structures. Nature *381* , 272.

Iqbal, S.A., Wallach, J.D., Khoury, M.J., Schully, S.D., and Ioannidis, J.P. (2016). Reproducible Research Practices and Transparency across the Biomedical Literature. PLoS biology *14* , e1002333.

Johnson, G.T., and Hertig, S. (2014). A guide to the visual analysis and communication of biomolecular structural data. Nat Rev Mol Cell Biol*15* , 690-698.

Kc, D.B. (2017). Recent advances in sequence-based protein structure prediction. Briefings in bioinformatics *18* , 1021-1032.

Khan, S., and Vihinen, M. (2009). Evaluation of accuracy and applicability of protein models: retrospective analysis of biological and biomedical predictions. In Silico Biol *9* , 307-331.

Kuhlman, B., and Bradley, P. (2019). Advances in protein structure prediction and design. Nat Rev Mol Cell Biol *20* , 681-697.

Laskowski, R.A., MacArthur, M.W., Moss, D.S., and Thornton, J.M. (1993). PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Cryst *26* , 283-291.

Lüthy, R., Bowie, J.U., and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. Nature *356* , 83-85.

Martí-Renom, M.A., Stuart, A.C., Fiser, A., Sánchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct *29* , 291-325.

Mura, C., McCrimmon, C.M., Vertrees, J., and Sawaya, M.R. (2010). An introduction to biomolecular graphics. PLoS computational biology*6* .

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera–a visualization system for exploratory research and analysis. J Comput Chem *25* , 1605-1612.

Pieper, U., Webb, B.M., Dong, G.Q., Schneidman-Duhovny, D., Fan, H., Kim, S.J., Khuri, N., Spill, Y.G., Weinkam, P., Hammel, M., *et al.* (2014). ModBase, a database of annotated comparative protein structure models and associated resources. Nucleic acids research*42* , D336-346.

Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. Journal of molecular biology *7* , 95-99.

Sarkans, U., Gostev, M., Athar, A., Behrangi, E., Melnichuk, O., Ali, A., Minguet, J., Rada, J.C., Snow, C., Tikhonov, A., *et al.*(2018). The BioStudies database-one stop shop for all data supporting a life sciences study. Nucleic acids research *46* , D1266-d1270.

Schwede, T., Sali, A., Honig, B., Levitt, M., Berman, H.M., Jones, D., Brenner, S.E., Burley, S.K., Das, R., Dokholyan, N.V., *et al.*(2009). Outcome of a workshop on applications of protein models in biomedical research. Structure *17* , 151-159.

Shapovalov, M.V., and Dunbrack, R.L., Jr. (2011). A smoothed backbone-dependent rotamer library for proteins derived from adaptive kernel density estimates and regressions. Structure *19* , 844-858.

Shen, M.Y., and Sali, A. (2006). Statistical potential for assessment and prediction of protein structures. Protein science : a publication of the Protein Society *15* , 2507-2524.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology *7* , 539.

Tan, T.W., Tong, J.C., Khan, A.M., de Silva, M., Lim, K.S., and Ranganathan, S. (2010). Advancing standards for bioinformatics activities: persistence, reproducibility, disambiguation and Minimum Information About a Bioinformatics investigation (MIABi). BMC Genomics*11 Suppl 4* , S27.

Vihinen, M. (2012). How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. BMC Genomics *13 Suppl 4* , S2.

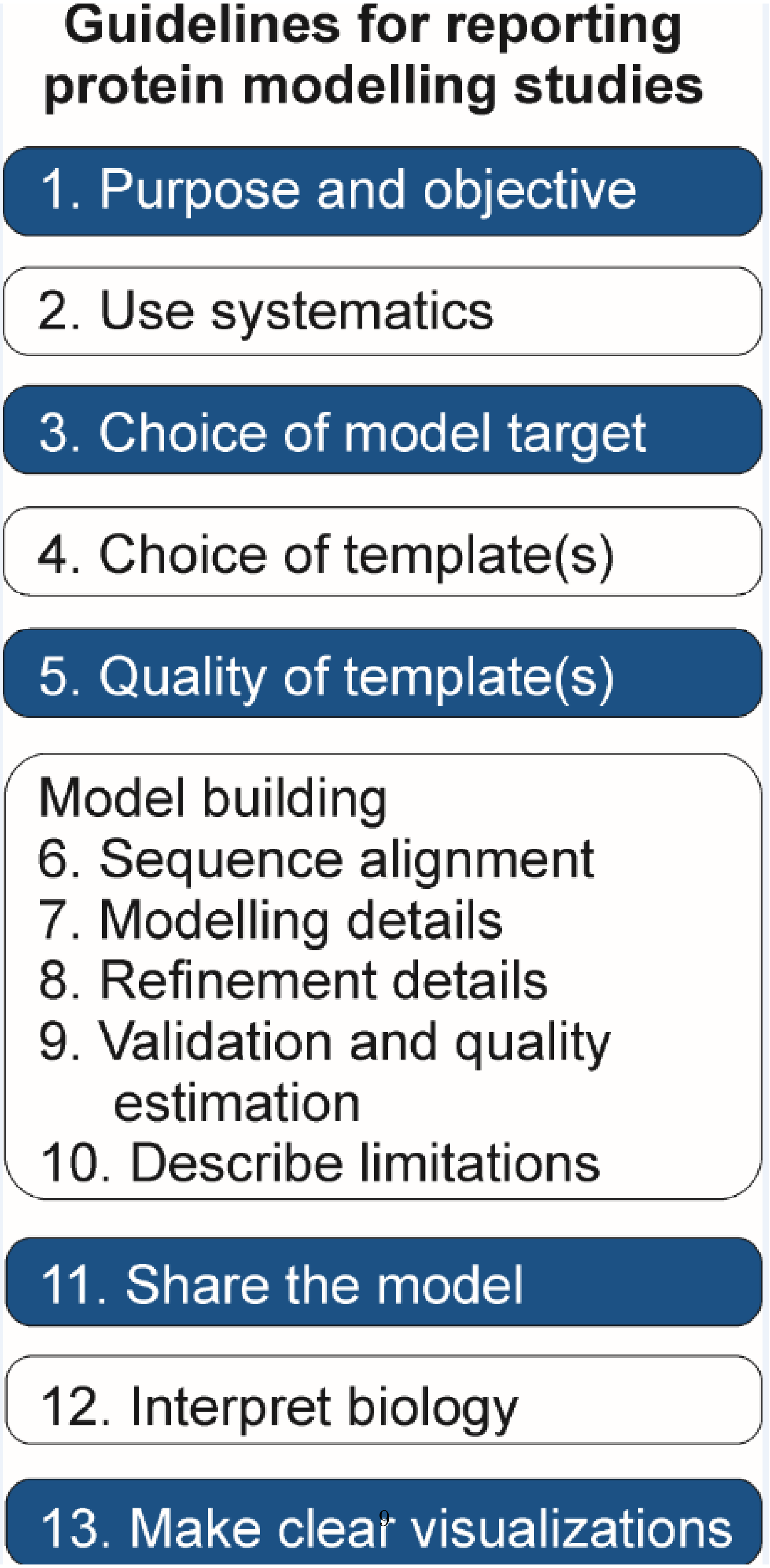
Vihinen, M. (2013). Guidelines for reporting and using prediction tools for genetic variation analysis. Human mutation *34* , 275-282.

Vihinen, M. (2014). Variation Ontology for annotation of variation effects and mechanisms. Genome research *24* , 356-364.

Vihinen, M. (2020). Guidelines for systematic reporting of sequence alignments. Biol Meth Protoc *bpaa001* .

Williams, C.J., Headd, J.J., Moriarty, N.W., Prisant, M.G., Videau, L.L., Deis, L.N., Verma, V., Keedy, D.A., Hintze, B.J., Chen, V.B.*, et al.* (2018). MolProbity: More and better reference data for improved all-atom structure validation. Protein science : a publication of the Protein Society *27* , 293-315.

Figure legend



Guidelines for reporting protein modelling studies

1. Purpose and objective

2. Use systematics

3. Choice of model target

4. Choice of template(s)

5. Quality of template(s)

Model building
6. Sequence alignment
7. Modelling details
8. Refinement details
9. Validation and quality estimation
10. Describe limitations

11. Share the model

12. Interpret biology

13. Make clear visualizations

Figure 1. Items and topics to be included into comprehensive description of molecular modelling study.